

# Deterministic Annealing, Constrained Clustering, and Optimization.

K. Rose<sup>†</sup>, E. Gurewitz\*, and G. C. Fox<sup>‡</sup>

Caltech Concurrent Computation Program

California Institute of Technology

## 1. Introduction

Clustering is usually formulated as an optimization problem by defining a cost function to be minimized. Most of these cost functions are nonconvex and have several local minima. Traditional techniques [1,2,3,4] are essentially decent algorithms, as the cost is reduced at each iteration. Therefore, they tend to get trapped in local minima.

Simulated annealing, or stochastic relaxation [5], is a known technique for avoiding local minima of nonconvex optimization problems. However, the process to escape local minima requires very slow schedules [6] which are not realistic for many practical applications.

In a previous work [7,8] we proposed the concept of deterministic annealing for the problem of clustering and vector quantization. Our approach is strongly motivated by the physical analogy and based on principals of statistical mechanics or information theory [9,10]. In this paper we extend our clustering method to the constraint clustering method. Sections 2 and 3 are a brief presentation of our clustering approach. Section 4 presents the "constraint clustering approach", and sections 5 and 6 are two examples for which this approach can be applied.

## 2. Clustering by Deterministic Annealing

In previous works [7,8] we have suggested a deterministic annealing approach to clustering and vector quantization. This approach is briefly summarized in this section and the following one.

The method uses a fuzzy formulation where each data point  $x$  is associated *in probability* with each cluster  $C_j$ . The cluster  $C_j$  is represented by its "cluster centroid"  $y_j$ . The energy (cost or distortion) of associating the data point  $x$  to the cluster  $C_j$  is  $d(x, y_j)$ . If the set  $\mathbf{Y} = \{y_j\}$  of the clusters representatives is given, the expected energy of the system is

$$\mathbf{E} = \sum_x \sum_j \mathbf{P}(x \in C_j) d(x, y_j) \quad (1)$$

where  $\mathbf{P}(x \in C_j)$  is the probability that the data point  $x$  belongs to the cluster  $C_j$ . Since we *do not* have any prior knowledge about the data probability distribution function, we apply the principle of maximum entropy. As is well known, the probability distribution which maximizes the entropy under the expectation constraint (1) is

$$\mathbf{P}(x \in C_j) = \frac{e^{-\beta d(x, y_j)}}{\mathbf{Z}_x}, \quad (2)$$

where  $\mathbf{Z}_x$  is the partition function  $\mathbf{Z}_x = \sum_j e^{-\beta d(x, y_j)}$ .

The Lagrange multiplier  $\beta$  is determined by a given value of  $\mathbf{E}$  in (1), and is inversely proportional to the "temperature". For a given set of cluster representatives  $\{y_j\}$ , it is assumed that the probabilities relating different data point to their clusters are independent. Hence the total partition function is  $\mathbf{Z} = \prod_x \mathbf{Z}_x$ , and the free energy  $\mathbf{F}$  is given by

$$\mathbf{F} = -\frac{1}{\beta} \log \mathbf{Z}(\mathbf{Y}) = \frac{1}{\beta} \sum_x \log \sum_j e^{-\beta d(x, y_j)} \quad (3).$$

<sup>†</sup> Department of Electrical and Computer Engineering, University of California Santa Barbara.

\* Physics Department, Nuclear Research Center Negev, Israel

<sup>‡</sup> University of Syracuse.

Instead of considering the association probability of a data point, we consider the probability of an entire instance. An instance is given by a set of representatives  $\mathbf{Y} = \{y_j\}$  and every data point is associated, in each instance, with one and only one representative. The energy of an instance is given by:

$$\mathbf{D}(\mathbf{Y}, \mathbf{V}) = \sum_x \sum_j v_{xj} d(x, y_j). \quad (4)$$

Where  $\mathbf{V} = \{v_{xj}\}$  is the set of associations,  $v_{xj} = 1$  if  $x \in C_j$ , and  $v_{xj} = 0$  for all the other clusters. In another study we are showing [11] that for a given  $\beta$  the most probable instance is the one with the minimal energy  $\mathbf{D}$ , and the set of the most probable representatives is the one with the minimal free energy. Thus, on one hand we have the energy  $\mathbf{D}$ , which is minimized to obtain the optimal, hard clustering solution for the training set. On the other hand we have the free energy  $\mathbf{F}$  which is minimized, at a given  $\beta$ , to obtain the set of the most probable representatives. Moreover, for  $\beta \rightarrow \infty$ , both  $\mathbf{D}$  and  $\mathbf{F}$  are minimized by the same set  $\mathbf{Y}$ .

### 3. Deterministic Annealing and Phase Transitions

The set  $\mathbf{Y}$  that optimizes the free energy satisfies:

$$\frac{\partial}{\partial y_j} \mathbf{F} = 0, \text{ hence, } \sum_x \mathbf{P}(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) = 0 \quad \forall j \quad (5)$$

Where we have used a shorthand notation for differentiation with respect to each component of  $y_j$  separately. After normalization (5) can be written as:

$$\left\langle \frac{\partial}{\partial y_j} d(x, y_j) \right\rangle_j = 0.$$

At  $\beta = 0$  the association probabilities are uniform and

$$\frac{\partial}{\partial y_j} \mathbf{F} = \frac{\partial}{\partial y_j} \sum_x d(x, y_j) = 0.$$

This is the centroid condition for the entire data set viewed as one cluster. As proved in [11] there is a unique solution to this equation. Thus at  $\beta = 0$  our effective function is convex, and all the representatives converge to the same point, which is the global minimum. We shall interpret identical representatives as representing the same cluster and consider the set  $\mathbf{Y}$  of representatives *without repetitions* as the set of natural clusters. Thus, at  $\beta = 0$  we have *one* natural cluster.

The single solution of (5) at  $\beta = 0$  will be a solution for all  $\beta > 0$ , but at some positive  $\beta$  it will change from stable solution (local minimum) to unstable solution (saddle point). At this point the system can gain energy by splitting into subgroups of representatives. Each subgroup defines a newly formed natural cluster. Hence, the system undergoes a phase transition. This process results in a natural hierarchy of clustering solution. Recalling that  $\beta$  is the Lagrange multiplier related to the average energy (distortion), we have clustering solutions at decreasing levels of average energy. It can be regarded as clustering at different scales.

An impotent example of a distortion measure is the  $\nu$ -th law  $d_\nu(x, y) = \sum_i |x(i) - y(i)|^\nu$ . The squared distance distortion,  $\nu = 2$ , is the most extensively used distortion measure. For this measure, we have shown [7,8] that the first transition occurs when  $(\mathbf{I} - 2\beta \mathbf{C}_{xx}) \mathbf{Y}_\lambda = 0$ , where  $\mathbf{C}_{xx}$  is the covariance matrix of the data,  $\mathbf{I}$  is the identity matrix, and  $\mathbf{Y}_\lambda$  is an eigenvector of  $\mathbf{C}_{xx}$ . The critical  $\beta$  is thus:  $\beta_c = (2\lambda_{max})^{-1}$ , where  $\lambda_{max}$  is the largest eigenvalue of  $\mathbf{C}_{xx}$ . Hence, the critical temperature is determined by the variance along the largest principal axis of the distortion. The split will be initialized in the direction of this principal axis.

### 4. Constrained Clustering

So far we did not put any constraint on the clusters' representatives. It was only implicitly assumed that there are enough representatives to enable phase transitions. Adding constraints to the deterministic annealing mechanism expands the variety of optimization problems which can be solved by this method.

There is a large family of optimization problems which may be viewed as looking for the optimal association between two sets, one set of variables and one set of fixed data. In the clustering problem these two sets are the set of data point and the set of clusters representatives. Optimal association, in this regard, means that an association cost function is defined, and has to be minimized.

Let us introduce constrains into our clustering formulation. In addition to the constraint of the average energy  $\langle D(\mathbf{Y}, \mathbf{V}) \rangle = \mathbf{E}$ , there is a constraint concerning only the cluster representatives  $\langle \mathbf{T}(\mathbf{Y}) \rangle = \mathbf{L}$ . The probability distribution which maximizes the entropy under these constraints is

$$\mathbf{P}(\mathbf{Y}, \mathbf{V}) = \frac{e^{-\beta D(\mathbf{Y}, \mathbf{V}) - \lambda \mathbf{T}(\mathbf{Y})}}{\sum_{\mathbf{Y}', \mathbf{V}'} e^{-\beta D(\mathbf{Y}', \mathbf{V}') - \lambda \mathbf{T}(\mathbf{Y}')}}.$$

By summing over all possible associations  $\mathbf{V}$  we obtain the marginal probability

$$\mathbf{P}(\mathbf{Y}) = \frac{e^{-\beta \mathbf{F}(\mathbf{Y}, \beta) - \lambda \mathbf{T}(\mathbf{Y})}}{\sum_{\mathbf{Y}'} e^{-\beta \mathbf{F}(\mathbf{Y}', \beta) - \lambda \mathbf{T}(\mathbf{Y}')}},$$

where  $\mathbf{F}$  is given in (3). The most probable  $\mathbf{Y}$  is the one which minimizes  $\beta \mathbf{F} + \lambda \mathbf{T}$ . It can be viewed as minimizing  $\mathbf{F}$  subjected to the constraint  $\mathbf{T}(\mathbf{Y}) = \mathbf{L}'$  ( $\mathbf{L}' = \mathbf{L}/\beta$ ). Hence, our procedure for constrained clustering is to minimize the free energy subjected to the appropriate constraint, or equivalently to minimize the Lagrangian:

$$\mathbf{F}' = \mathbf{F} + q\mathbf{T} \quad \text{where } q = \lambda/\beta \quad (6)$$

## 5. Mass-Constrained Clustering

In our clustering method the location of the clusters' representatives depends on their number. Therefore, we introduce a constraint taking into account the mass, or the population of a natural cluster.

Let  $\lambda_k$  denote the multiplicity of representatives in the  $k$ -th cluster. Then  $\mathbf{Z}_x = \sum_j e^{-\beta d(x, y_j)}$ , is rewritten as  $\mathbf{Z}_x = \sum_k \lambda_k e^{-\beta d(x, y_k)}$ , the free energy (3) is now

$$\mathbf{F} = -\frac{1}{\beta} \sum_x \log \mathbf{Z}_x = -\frac{1}{\beta} \sum_x \sum_k \lambda_k e^{-\beta d(x, y_k)}.$$

The free energy is to be minimized under the constraint of a fixed number of representatives. The Lagrangian to be minimized is thus

$$\mathbf{F}' = \mathbf{F} + q \left( \sum_k \lambda_k - M \right).$$

We do not require  $\lambda_k$  to be integers, and  $M$  is the total mass of the natural clusters. The set  $\{\lambda_k\}$  which minimize  $\mathbf{F}'$  satisfies

$$\frac{\partial}{\partial \lambda_k} \mathbf{F}' = -\frac{1}{\beta} \sum_x \frac{e^{-\beta d(x, y_k)}}{\mathbf{Z}_x} + q = 0,$$

which yields

$$q\beta = \sum_x \frac{e^{-\beta d(x, y_k)}}{\mathbf{Z}_x} \quad \forall k.$$

Multiplying by the appropriate  $\lambda_k$  and summing over all natural clusters

$$\sum_k \lambda_k q\beta = \sum_k \lambda_k \sum_x \frac{e^{-\beta d(x, y_k)}}{\mathbf{Z}_x} = \sum_x \sum_k \mathbf{P}(x \in C_k),$$

which by the mass constraint yields  $q\beta = N/M = 1/\mu$ , where  $N$  is the number of data points and  $\mu$  is the mass of a data point. From the last equation we get

$$\lambda_j = \mu \sum_x \frac{\lambda_j e^{-\beta d(x, y_j)}}{Z_x} = \sum_x \mu P(x \in C_j).$$

Note that although  $\mu$  is constant, it could be made dependent on  $x$  to generalize the method to cases where the given data points do not have equal weights.

In the mass-constrained formulation the process is independent of the number of representatives, as long as this number is greater than the number of natural clusters. Let the natural clusters be represented by the clusters centroids  $\{y_j\}$  and masses  $\{\lambda_j\}$ . The  $j$ -th natural cluster  $C_j$  is represented by its centroid  $y_j$  and its mass  $\lambda_j$ . If this natural cluster is represented by more than one centroid, i.e.  $m_j$  and  $C_j = \cup_k C_j^{(k)}$ , it is represented by the two sets  $\{y_j^{(k)}\}$  and  $\{\lambda_j^{(k)}\}$ ,  $k = 1, \dots, m_j$  and  $\sum_{k=1}^{m_j} \lambda_j^{(k)} = \lambda_j$ . From the definition of  $Z_x$  it is clear that  $Z_x$  is invariant to any division of the natural cluster into clusters. Furthermore, the probability of associating a data point  $x$  to a natural cluster does not change

$$\sum_k P(x \in C_j^{(k)}) = \sum_k \frac{\lambda_j^{(k)} e^{-\beta d(x, y_j)}}{Z_x} = \lambda_j \frac{e^{-\beta d(x, y_j)}}{Z_x} = P(x \in C_j).$$

It is therefore clear that these representatives will satisfy (5), and will thus be obtained by our method regardless of the multiplicity  $m_j$  or the mass division  $\{\lambda_j^{(k)}\}$ .

## 6. The Travelling Salesman Problem

We shall derive here the Elastic Net [12,13] approach to the travelling salesman problem (TSP) as an application of our constraint clustering procedure.

In the deterministic annealing clustering algorithm, if we throw in enough representatives and let  $\beta \rightarrow \infty$ , then each data point becomes a natural cluster. This can be viewed as a process of mapping a set of data points to a set of representatives.

The TSP is: Given a set of data points (cities), find the shortest close path which passes through all of them. Actually, like in the Elastic Net approach, instead of minimizing the length we shall minimize the sum  $L$  of the squared distances between consecutive cities on the path ( $L$ ). Hence, we add the constraint to the free energy to obtain the Lagrangian (6)

$$F' = F + \lambda \left( \sum_{k=1}^n |y_k - y_{k-1}|^2 - L \right), \quad y_0 = y_n.$$

The optimal  $Y$  must satisfy the condition  $\frac{\partial}{\partial y_j} F' = 0$ , for all the  $j$ 's, which by (5) yields

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) + \lambda(2y_j - y_{j+1} - y_{j-1}) = 0.$$

If we choose the square distance as our clustering distortion measure  $d(x, y)$ , we obtain the EN formulation for the optimum.

$$\sum_x P(x \in C_j) (y_j - x) + \lambda(2y_j - y_{j+1} - y_{j-1}) = 0.$$

This equation depends on  $\beta$  through the association probabilities  $P$ . When  $\beta \rightarrow \infty$ , each representative converge to one city.

An important question at this point is how  $\lambda$  should be varied with  $\beta$ . In [12] the relation was  $\lambda \propto 1/\sqrt{\beta}$ , while in [14] it was kept constant.

The procedure suggested here is as follows: a) For a given  $\beta$ , gradually increase  $L$  and optimize, until  $L$  reaches some appropriate value below the free (of constraint) length. b) Keeping  $L$  constant, update  $\beta$  and optimize, return to (a). Such an approach can be implemented directly using methods for nonlinear optimization, like for example Generalized Hopfield Networks.

It is more convenient to control the Lagrange multiplier  $\lambda$  rather than the length  $L$ . It was shown [11] that one can use the following approximation:

$$\Delta\lambda(\beta) \approx -\frac{\Delta\beta}{\beta} \left( \frac{\Delta E}{\Delta L} + \lambda \right),$$

where  $\Delta E/\Delta L$  may be estimated using the last two iterations in  $\lambda$ , before the moment to update  $\beta$ .

### Concluding Remarks

Constrained clustering can be used to improve the deterministic annealing method for clustering. Moreover, it allows applying annealing to various other optimization problems which have a data association aspect, and can be formulated as constrained clustering. Annealing is obtained by gradually making the association less fuzzy, and helps avoiding local minima.

### References

- [1] G. Ball and D. Hall, "a clustering technique for summarizing multivariate data", *Behavioral Science*, vol. 12, pp. 153-155, 1967.
- [2] J. C. Dunn, "a fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *J. Cybrn*, vol. 3, pp. 32-57, 1974.
- [3] J. C. Bezdek, "A convergence theorem for the fuzzy isodata clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, pp. 1-8, 1980.
- [4] Y. Linde, A. Buzo, and R. M. Gray, "an algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84-95, 1980.
- [5] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671-680, 1983.
- [6] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, pp. 721-741, 1984.
- [7] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transition in clustering," *Physical Review Letters*, vol. 65, pp. 945-948, 1990.
- [8] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing," to be published.
- [9] E. T. Jaynes, "Information theory and statistical mechanics," in *Papers on probability, statistics and statistical physics* (R. D. Rosenkrantz, ed.), Dordrecht, The Netherlands: Kluwer Academic Publishers, 1989.
- [10] C. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana, Ill.: The University of Illinois Press, 1949.
- [11] K. Rose, *Deterministic annealing, clustering and optimization*. Ph.D. Thesis, California Institute of Technology, Pasadena California, 1991.
- [12] R. Durbin and D. Willshaw, "An analogue approach to the travelling salesman problem using an elastic net method," *Nature*, vol. 326, pp. 689-691, 1987.

- [13] R. Durbin R. Szeliski, and A. L. Yuille, "An analysis of the elastic net approach to the travelling salesman problem," *Neural Computation*, vol. 1, pp. 348-358, 1989.
- [14] A. L. Yuille, "generalized deformable models, statistical physics, and matching problems," *Neural Computation*, vol. 2, pp. 1-24, 1990.

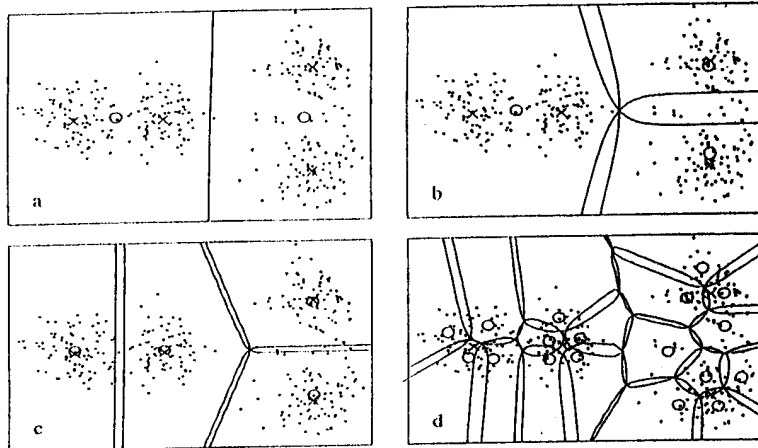


FIG. 1 Clustering at different phases. The data is generated from four Gaussian distribution centered at the locations marked by x. The calculated centers are marked by o.

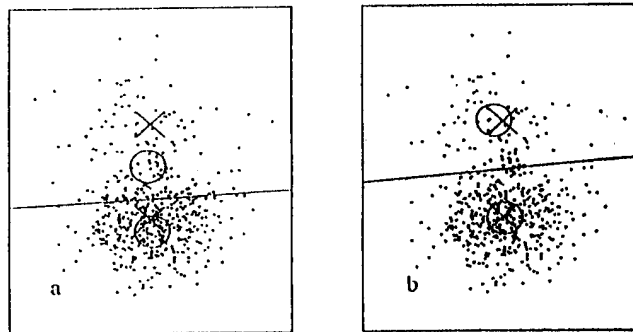


FIG. 2 The effect of cluster mass (population) at intermediate  $\beta$ . The data is sampled from two normal distributions whose centers are marked by x. The computed representatives are marked by o.

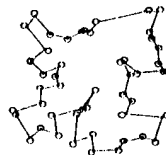


FIG. 3 A constrained clustering result for a 50 cities TSP.